Toward automated infrared spectral analysis in community drug checking

Lea Gozdzialski, Abby Hutchison, Bruce Wallace, Chris Gill, Dennis Hore

2023

Received: 16 March 2023 | Revised: 13 May 2023 | Accepted: 16 May 2023

DOI: 10.1002/dta.3520

RESEARCH ARTICLE

WILEY

Toward automated infrared spectral analysis in community drug checking

Lea Gozdzialski[1] | Abby Hutchison[2,3] | Bruce Wallace[2,4] | Chris Gill[1,2,5,6,7] | Dennis Hore[2,8]

[1]Department of Chemistry, University of Victoria, Victoria, British Columbia, Canada

[2]Canadian Institute for Substance Use Research, University of Victoria, Victoria, British Columbia, Canada

[3]School of Public Health and Social Policy, University of Victoria, Victoria, British Columbia, Canada

[4]School of Social Work, University of Victoria, Victoria, British Columbia, Canada

[5]Department of Chemistry, Applied Environmental Research Laboratories (AERL), Vancouver Island University, Nanaimo, British Columbia, Canada

[6]Department of Chemistry, Simon Fraser University, Burnaby, British Columbia, Canada

[7]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, Washington, USA

[8]Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada

Correspondence
Dennis Hore, Department of Chemistry, University of Victoria, Victoria, British Columbia V8W 3V6, Canada.
Email: dkhore@uvic.ca

Funding information
Health Canada's Substance Use and Addictions Program (SUAP); Vancouver Foundation

Abstract

The body of knowledge surrounding infrared spectral analysis of drug mixtures continues to grow alongside the physical expansion of drug checking services. Technicians trained in the analysis of spectroscopic data are essential for reasons that go beyond the accuracy of the analytical results. Significant barriers faced by people who use drugs in engaging with drug checking services include the speed and accuracy of the results, and the availability and accessibility of the service. These barriers can be overcome by the automation of interpretations. A random forest model for the detection of two compounds, MDA and fluorofentanyl, was trained and optimized with drug samples acquired at a community drug checking site. This resulted in a 79% true positive and 100% true negative rate for MDA, and 61% true positive and 97% true negative rate for fluorofentanyl. The trained models were applied to selected drug samples to demonstrate a proposed workflow for interpreting and validating model predictions. The detection of MDA was demonstrated on three mixtures: (1) MDMA and MDA, (2) MDA and dimethylsulfone, and (3) fentanyl, etizolam, and benzocaine. The classification of fluorofentanyl was applied to a drug mixture containing fentanyl, fluorofentanyl, 4-anilino-*N*-phenethylpiperidine, caffeine, and mannitol. Feature importance was calculated using shapely additive explanations to better explain the model predictions and k-nearest neighbors was used for visual comparison to labelled training data. This is a step toward building appropriate trust in computer-assisted interpretations in order to promote their use in a harm reduction context.

KEYWORDS
drug checking, explainable AI, infrared spectroscopy, random forest classification, SHAP

1 | INTRODUCTION

Recently the term "drug checking technician" has emerged. This term typically refers to people who are trained to perform drug analysis using a range of on-site methods and instruments, such as Fourier transform infrared (FTIR) spectrometers, colorimetric reagent testing, and immunoassay and test strips, and consolidate information to make conclusions about the substance composition.[1,2] The number

of drug checking technicians is rapidly multiplying as drug checking services expand both in the community, safe consumption sites, and at festivals.[2–4] Within communities where testing is done with FTIR, knowledge surrounding the spectral analysis of common drug mixtures within the field continues to grow alongside this expansion. Technicians learn the spectral patterns to look for when testing common drug mixtures, usually with the help of spectral libraries and software. For example, detecting a low concentration of fentanyl in a mixture of bulking agents (commonly referred to as cuts and buffs) is typically challenging for any basic library searching scheme. However, a trained drug checking technician can immediately recognize areas in the spectrum where fentanyl has minimal overlap with the cutting agents (e.g. caffeine or mannitol), manually find evidence of fentanyl in the spectrum, and consider the findings along with contextual evidence (e.g., suspected substance, appearance, and anecdotal evidence if the substance has been used). At the same time, there is concern that this level of subjective interpretation in analyzing FTIR spectra may lead to misleading results.[5] Such complexity is recognized as a significant barrier to implementing drug checking services, and indirectly affects aspects such as speed, accuracy, availability, and accessibility of drug checking.[6,7] These areas (speed, accuracy, etc.) have been recognized as significant factors affecting the willingness of people who use drugs (PWUD) to engage with drug checking.[7] It is acknowledged that current drug checking technologies, including the associated software for data analysis and interpretation, are still under development.[6]

Many of the advantages, and challenges, of drug checking with FTIR are inherent to the hardware and underlying technology (e.g., ease-of-use, non-destructive, limited capabilities for low concentration components, and complex mixtures).[1,8] However, some areas of the implementation, particularly in the interpretation of the IR spectra, could possibly be improved through software. Machine learning (ML) broadly refers to a group of algorithms that reveal patterns in a set of data, connects those patterns to a meaningful result, and uses that relationship to predict future unknown data.[9] ML has been used for guiding spectroscopic interpretation for many years and the literature is rich with examples, methods, and proposed workflows.[9–13] For example, automation can speed up spectral analysis, alleviate the requirement of an experienced technician, and can offer a greater degree of consistency, accuracy, and precision in the reported results. However, common barriers to implementing ML include the requirement of a large quantity of high-quality data with known labels for building ML models, time and expertise required for the testing and implementation of such models, and protocols for on-going validation. Applying ML to drug checking faces additional barriers, namely, obtaining exemptions for controlled substances, and coordinating off-site confirmatory testing.

Despite the many advances in ML for spectroscopic interpretation, as well as the growing body of past drug checking data (e.g., IR spectra with associated interpretations/labels), spectroscopy-based drug checking has surprisingly not been fully automated. To date, most drug checking services rely on a human decision-maker (i.e., drug checking technician or harm reduction worker) for the spectral interpretation and final result,[1,5,14] even in cases where ML models exist to support that interpretation (e.g., quantification of fentanyl[15]). Complete automation risks the exclusion of experiential community knowledge, and the erasure of opportunities for the co-production of knowledge that combines technology and this experiential knowledge.[6,16] For instance, higher levels of satisfaction have been associated with drug checking services that operate in the contexts of clear communications and transparency.[4] Curiosity in the instruments and analytical process used in drug checking has also been noted to drive engagement of PWUD with drug checking service.[16,17] Overall, these studies motivate the consideration of contextual information and the value of curiosity when developing tools to aid in spectral interpretation.

Explainable artificial intelligence (XAI) guides explanations of predictions provided by ML models, and explores methods to present such explanations (e.g., visualizations, text, and interactive tools).[18–22] Some methods include exposing feature importance, that is, what part of the input was most influential in the prediction, and presenting "learn-by-example" cases. For example, technicians familiar with the relationship between spectral features and the presence or absence of certain compounds want to see such evidence. XAI also addresses the trust in and transparency of ML models.[21,23] In general, the pursuit of model explanations is motivated by three main purposes: model validation, model debugging, and knowledge discovery.[23,24]

This work uses a supervised machine learning algorithm trained on IR spectral data with associated paper spray mass spectrometry (PS-MS) results to predict the presence of target compounds in unknown samples. XAI methods that expose the reasoning behind classification decisions are presented to (a) connect with current practices for interpreting drug checking data, (b) allow for on-going human-in-the-loop interference for improvements and quality assurance of ML models, and (c) promote continuous knowledge production within drug checking services both for technicians and people engaging with the service.

2 | METHODS

2.1 | Data acquisition

Infrared spectra ($n = 7091$) used in this study were acquired between November 2020 and November 2022 through our service, Substance, the Vancouver Island Drug Checking Project,[25] located in Victoria, British Columbia. A portable FTIR with a 45° diamond ATR element (Agilent 4500a) was used. Spectra were acquired with 32 averages and an effective resolution of 4 cm^{-1}. Samples are received in various forms, with majority of substances tested as powders. This subset of IR spectra was chosen for building and evaluating classification models such that the same drug sample was also analyzed using paper spray mass spectrometry (PS-MS), for its ability to more unambiguously report on the presence or absence of particular trace actives.[26–28] Details of the PS-MS method as used for drug checking,

including providing quantitative information, has been previously described in detail.[26,27,29,30] The overall composition for each sample was ultimately determined through IR analysis performed by a trained technician, using tools such as library matching, together with PS-MS in a point-of-care setting. It is noted that some of the components may be missed when they are below the LOD of IR-based methods, and not distinguishable on PS-MS.

2.2 | Random forest (RF)

An RF classifier was used for binary classification. RF is an ensemble (voting) classifier that uses a series of classification trees, each built on a random subset of input features. The RF classifiers were implemented using the *scikit-learn* package in python.[31] Various combinations of preprocessing and hyperparameters were first explored using 3-fold cross validation. A random search cross validation procedure, implemented using *scikit-learn*, was used for the initial narrowing of the optimal preprocessing and hyperparameter space to further pursue with more comprehensive optimizations. The F1 score, a harmonic mean of the precision and recall, was used to evaluate these combinations. For the MDA model, a more comprehensive grid of hyperparameters was pursued using standard normal variate (SNV) preprocessing to maximize the F1 score. Similarly, the fluorofentanyl model was further optimized to maximize the F1 score during cross validation using min-max and second derivative spectral preprocessing. The grid of hyperparameters and preprocessing, as well as the performance metrics for both models, are shown in Tables S1–S4. To simplify the model, a subset of the most important features, as calculated by the Gini index, was used to achieve similar performance with the RF. $n = 100$ and $n = 20$ features were chosen for the MDA and fluorofentanyl model, respectively, and these models were used for validation and additional analysis of selected drug mixtures.

2.3 | K-nearest neighbors (KNN)

KNN[32] was used to find the closest matching spectra to generate examples for visualization purposes. Two models, one for positive samples and one for negative samples, were built with the respective training spectra. All spectra were preprocessed and truncated according to the optimized RF classification model. KNN was implemented using *scikit-learn* where n_neighbors = 2 and metric = "euclidean."[31]

2.4 | Shapely additive explanations (SHAP)

The SHAP method was used to estimate the contribution of each input feature to the final prediction, therefore attempting to generate an explanation to an end-user regarding the model decision.[23,33] For our models, Kernel SHAP was implemented, a popular model-agnostic method, via the python package *shap*.[23,33]

2.5 | Sample selection

Two different target compounds were selected for building two classification models. The first model aimed to detect 3,4-methylenedioxyamphetamine (MDA), and the second was based on fluorofentanyl detection. These compounds were chosen based on their prevalence in and relevance to the local drug supply, and the fact that they represent simple (MDA) and more challenging (fluorofentanyl) classification problems. Furthermore, they have different potencies, with MDA usually appearing in high concentrations (or pure form) and fluorofentanyl typically being significantly cut. The MDA classification model was trained with IR absorption spectra of drug samples received at the drug checking service ($n = 4963$). Each spectrum was labeled based on whether it represented a sample with or without MDA as 0 (not present/"negative," $n = 4804$) or 1 (present/"positive" $n = 159$), as previously determined by PS–MS, regardless of the other compounds present in the sample. For example, if a drug sample was determined to contain both 3,4-methylenedioxymethamphetamine (MDMA) and MDA through secondary testing with PS–MS, then it is labelled as "1." An external test set ($n = 2060$ without MDA and $n = 68$ with MDA) was used for validating the final, optimized classification model. Similarly, the RF model to detect fluorofentanyl was trained with a subset of IR spectra of samples determined to be within the category of opioid or "down," received at the drug checking service ($n = 2575$). Each sample was labeled based on whether it represented a sample with or without fluorofentanyl as 0 (not present/"negative," $n = 2202$) or 1 (present/"positive," $n = 373$), as previously determined by PS–MS, regardless of the other compounds present in the sample. Although MS is unable to differentiate between fluorofentanyl isomers, the IR spectral features were consistent with para-fluorofentanyl, hereafter referred to simply as fluorofentanyl. An external test set ($n = 551$ without fluorofentanyl and $n = 93$ with fluorofentanyl) was used to validate the trained model.

For the application of XAI methods KNN and SHAP, three test mixtures were chosen for prediction with the trained MDA model, representing the cases of correct positive prediction (5% MDA by weight in MDMA), incorrect negative prediction (54% MDA in dimethylsulfone), and correct negative prediction (an opioid mixture containing fentanyl, benzocaine, and etizolam). For the fluorofentanyl example, a correct positive prediction (6% fluorofentanyl in a mixture containing fentanyl HCl, 4-anilino-*N*-phenethylpiperidine commonly known as ANPP, caffeine, and mannitol) was chosen as an opportunity to demonstrate the ability of XAI to highlight subtle spectral features that have contributed to classification.

3 | RESULTS AND DISCUSSION

3.1 | Model performance and optimization

The initial optimization of the RF model to detect MDA is illustrated in Figure 1a, where the color of the grid relates to the F1 score of

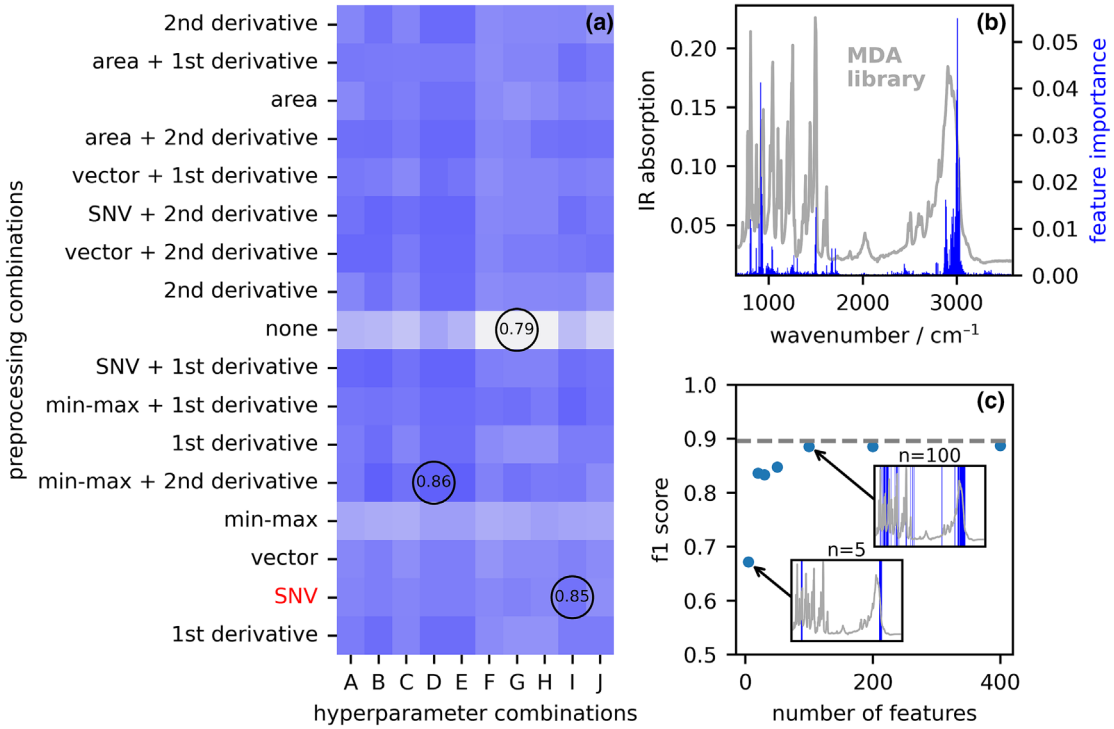

FIGURE 1 (a) Matrix representing combinations of hyperparameters and spectral preprocessing for optimizing performance of the RF model based on the F1 score. (b) Feature importance calculated within the RF model. Notably, the most important features identified correlate well to strong features in pure MDA. (c) F1 score on the test set confined to the *n* most important features as calculated from the base RF model.

cross validation. The lowest F1 score of cross validation was calculated to be 0.79 with no preprocessing and the highest F1 score was calculated to be 0.86 with min-max normalization and second derivative preprocessing. SNV preprocessing was chosen, however, as the preprocessing of choice with a similarly high F1 score of cross validation (0.85). This decision was made because spectra with less alteration are more likely to align with technicians' understanding of IR absorption and ultimately contribute to the interpretability of the model and *post-hoc* visualizations. To understand the model's decision-making process in a general, or "global," sense the most influential features as determined by the Gini index of the RF model are calculated and shown in Figure 1b. The most important features align well with strong modes shown in the library entry for MDA, which is overlaid in gray. The final classification model used in subsequent analysis uses the top 100 most important features, discarding features with minimal influence on the prediction of MDA (Figure 1c). The resulting confusion matrix for the validation set is shown in Figure 2a. The precision was calculated on the external test set as 100% and the recall was determined to be 81%. The F1 score was calculated to be 0.9. The receiver operating characteristic (ROC) curve for the external test set is presented in Figure 2b. The area under the curve, which can be derived from the ROC, is a general performance measurement of how well two classes can be separated on a scale from 0 to 1. Here, an area under the ROC (AUROC) of 0.99 was obtained which demonstrates that there is excellent distinction between samples with MDA and samples without MDA.

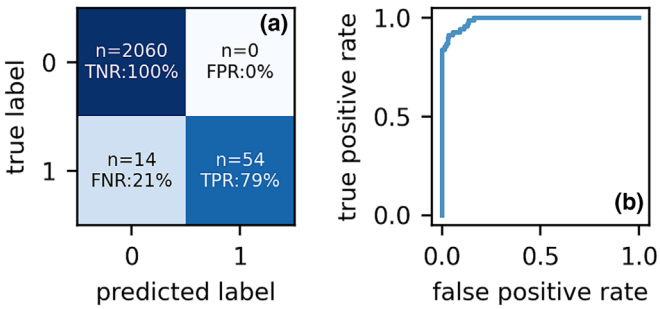

FIGURE 2 (a) Confusion matrix for external test set ($n = 2128$) using the optimized classification model from Figure 1. (b) Receiver operating characteristic curve (ROC) for the external test set, demonstrating the trade-off between true positives rates and false positive rates. Each point along the graph represents a varied decision threshold for classification.

In optimizing the second RF model for fluorofentanyl detection (Figure 3a), the second derivative preprocessing was found to be necessary to resolve fluorfentanyl's sharp, highly overlapped and low intensity features within a crowded fingerprint region.[34] The feature importance inherent to the RF model, as calculated by the Gini index, revealed that the most important features align with strong features in the library fluorofentanyl spectrum in Figure 3b. Again, to simplify the model, a subset of these features was used to achieve similar, and in some cases, greater performance, than using the entire spectrum. This is shown in Figure 3c. $n = 20$ features were chosen and that

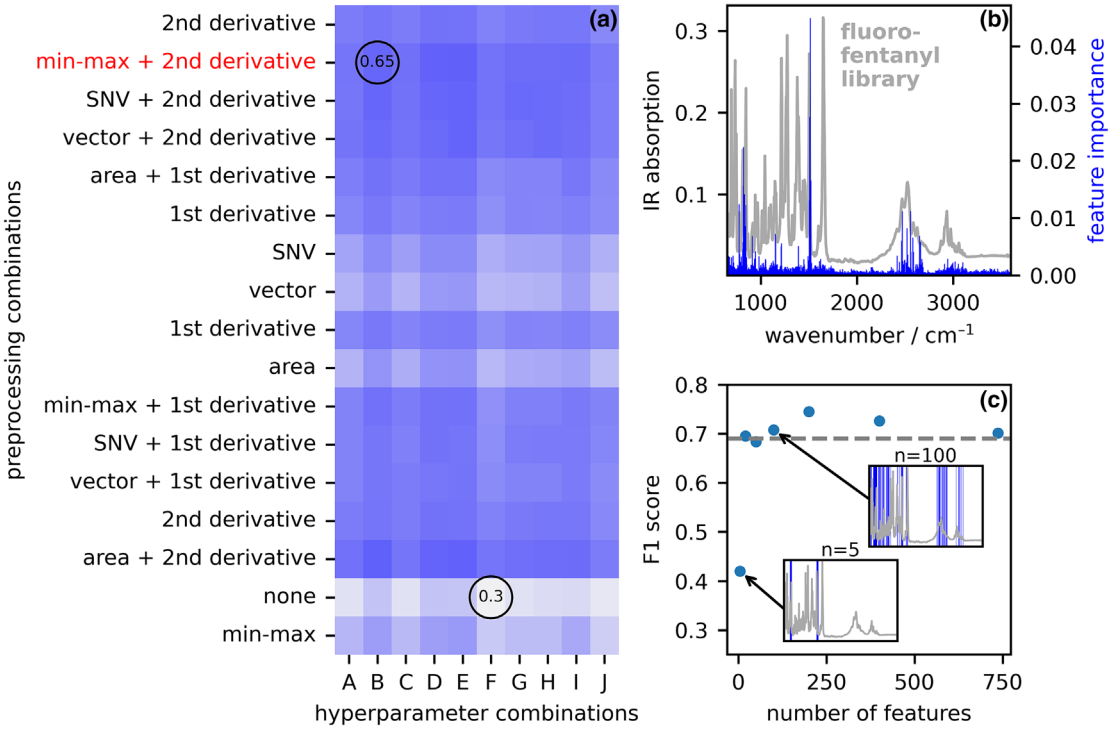

FIGURE 3 (a) Matrix representing combinations of hyperparameters and spectral preprocessing for optimizing performance of the RF model. The metric considered is the F1 score. (b) Feature importance calculated within the RF model. The few important features found correlate well to strong features in pure fluorofentanyl. Notably, many strong features of fluorofentanyl have minimal importance for the prediction. (c) F1 score on the test set for *n* most important features as calculated from the base RF model.

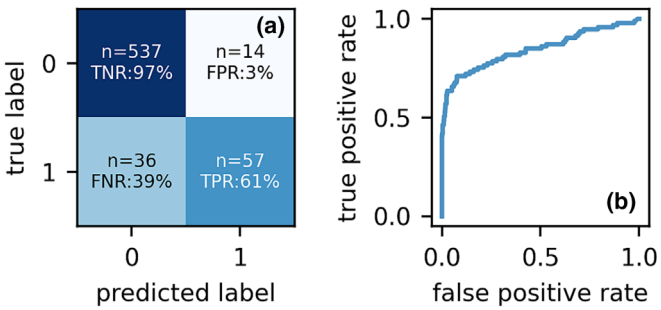

FIGURE 4 (a) Confusion matrix for external test set ($n = 644$). (b) ROC curve (receiver operating characteristic curve) demonstrating the trade-off between true positives rates and false positive rates. Each point on the graph represents a varied decision threshold for classification.

model was used for validation and additional analysis. The resulting confusion matrix is shown in Figure 4a. The precision and recall of the model was calculated as 80% and 61%, respectively, for the test set. The AUROC curve (Figure 4b) was calculated as 0.85, revealing the greater ambiguity in isolating unique spectral features between samples with and without fluorofentanyl.

It is well known that spectroscopy-based techniques such as FTIR and Raman lack sensitivity, yet are attractive for community drug checking because they are easy-to-use, robust, portable, and lower-cost. The trade-off of this advantage is a higher limit of detection and therefore, false negatives are inevitable when low concentration components exist in the drug market. There are no formal acceptance criteria established to indicate whether a model is suitable for deployment in the context of community drug checking, and in other applications it is often stated to be "fit for purpose,"[35] acknowledging associated risks of reporting on analytical results with uncertainty. In general, these classification models are expected to have high precision (confidence in positive hits if enough spectral evidence is present) but low recall (less confidence in negative hits), as seen before in drug checking applications.[30] For example, manual FTIR interpretation using spectral matching software has been shown to result in a similar outcome, where false negatives are far more prevalent than false positives due to the relatively high limit of detection of FTIR.[14,36,37] During validation, when the MDA model predicted that MDA was present in a sample, it was correct every time. However 19% were predicted as false negatives. A similar situation was found for the detection of fluorofentanyl in opioid samples. In this case, where the median concentration of fluorofentanyl in the training set was low (2.1 w/w%), IR spectroscopy approaches its limit of detection. Fourteen (3%) false positive predictions were made using the optimized model, however, fluorofentanyl was correctly detected in only about half of the positive test cases. The ROC curves shown in Figure 2b and Figure 5b demonstrate that, if one is willing to accept an increased risk of false positives as a trade-off for improved true positive detection, such a compromise can be considered.

It is noted that in the dataset presented here, as in most drug checking datasets, the populations of the classes are typically small, significantly unbalanced, and prone to label error as the overall composition is rarely known with absolute certainty. Determining what a suitable dataset is to begin the pursuit of ML does not have a straightforward answer. This depends on the problem at hand, the quality and complexity of the data, and the algorithm of choice.[38] RF classification was explored in this application because it minimizes overfitting due to its iterative bagging and voting,[39] performs an implicit feature selection by only using features that are most influential on reducing classification error,[40] and therefore is well suited to real-world datasets with some degree of label error.[39,40] Variations of the original RF model, such as balanced RF, were also used to overcome some limitations with an extremely imbalanced training set.[39] The initial model optimization, evaluation of performance, and interrogation of the relevance of features learned by the model are important steps in determining the suitability of a drug checking dataset for ML. Metrics such as precision and recall will guide whether a particular dataset is suitable however could possibly be further improved by using more training data, addressing label errors, data augmentation, data fusion or exploring complex neural network architectures. Eventually, for many target compounds, the error will be mostly attributed to the fact that both the training and test sets, and future samples received at the drug checking service, will contain drugs that are present below the limit of detection. Recognizing these limitations is important when considering the expansion of ML models for a range of compounds.

3.2 | Generating model explanations

Three "unknown" spectra from the test set for the classification of MDA are shown in Figure 5a–c to demonstrate the outputs from SHAP and KNN. First, the SHAP values were calculated for each test instance (Figure 5d–f). Second, KNN was used to retrieve factual and counterfactual explanatory cases (Figure 5g–i). The top four nearest neighbors from the training set, with known composition, are shown. Two nearest neighbors are from the positive class (MDA present, traces shown in red) and two nearest neighbors are from the negative class (no MDA present, trace shown in blue) with their corresponding correlation to the query spectrum. Together, these features aim to explain why a model might have predicted one class vs the other for a particular unknown sample given the IR absorption data.

The first case represents a mixture of MDMA and MDA, where MDMA is the major component, and there is minor contribution to the overall IR spectrum from MDA. The finalized model predicted the presence of MDA, with a probability of 100%. This suggests the

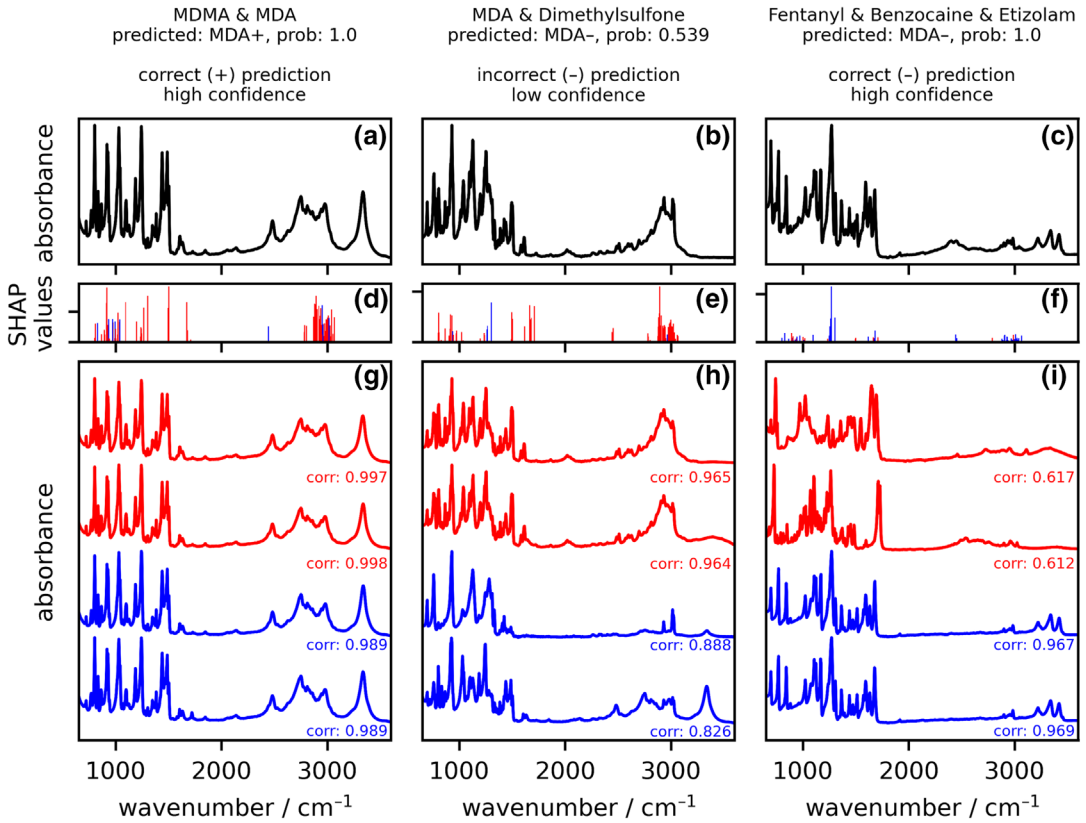

FIGURE 5 Various case examples using a combination of SHAP and KNN to aid in summarizing the RF classification results. Each column represents a different test case, with the "unknown" query spectrum shown in black (a–c). SHAP values of features that are found to contribute to a positive prediction are shown in red, and those that contribute to a negative prediction are shown in blue (d–f). Nearest neighbors traces that have MDA present are shown in red, and traces that do not have MDA present are shown in blue (g–i).

model is highly confident in the prediction, and the SHAP values draw attention to the features in the query spectrum that most contributed to the higher confidence (red) and features that possibly did not align with the presence of MDA (blue). As a feature of XAI, this prediction is supported by the observation that intensity in these regions directly corresponds to MDA vibrational modes. The four nearest neighbors demonstrate that both samples with MDA (+) and without MDA (−) in the database have a very high correlation with the query spectrum (Pearson's correlation 0.99). With these nearest neighbors and SHAP values together, technicians may try to extract evidence that the query spectrum does in fact have features consistent with the presence of MDA.

The second case is a mixture of MDA and dimethylsulfone. The model incorrectly predicted that there is no MDA present, where the prediction probability of 55% suggests uncertainty about the presence or absence of MDA. The SHAP values reveal that there are some features that do support the presence of MDA, however it was not significant enough to result in a positive detection. The four nearest neighbors reveal that the positive nearest neighbors, both which are samples with MDA and dimethylsulfone, have much higher correlation with the query spectrum (0.96) than the nearest neighbors from the negative class (0.82–0.88). Upon further inspection, there are in fact features consistent with the presence of MDA in our query spectrum and may disagree with the model here. This brings attention to the fact that perhaps this drug combination is less frequent and was not well represented in the training.

The final case is a mixture that includes fentanyl, benzocaine and etizolam. The model predicted that there was no MDA in this sample, with a prediction probability of 100%. The SHAP values reveal that almost no features of the query spectrum contributed to a positive prediction and highlights features that strongly suggest the absence of MDA. The nearest neighbors from the negative class have very high correlation to the query spectrum, further instilling confidence in this prediction. The nearest neighbors from the positive class have very poor correlation scores (0.61) in contrast to what was observed in the two previous examples (Figure 5g,h). Here there were no MDA positive nearest neighbors with a similar IR spectrum, further supporting that MDA in such a drug mixture was unlikely.

The same two methods for facilitating model explanations described in the previous examples were explored for fluorofentanyl classification. In this case, however, there were additional challenges because (1) the final model had poorer performance as a result of the low concentrations of fluorofentanyl and higher complexity of opioid drug mixtures, and subsequently, (2) greater spectral manipulation was required for optimal separation between the two classes (second derivative) and therefore less intuitive to use for visualizing spectral features. However, since the presence of fluorofentanyl was mostly determined from sharp features in the fingerprint region, there is a benefit from having to visualize and investigate fewer ($n = 20$) features. The implementation was demonstrated using a sample that contains fentanyl, fluorofentanyl, caffeine, and mannitol (Figure 6). Here, the fluorofentanyl model correctly predicted the

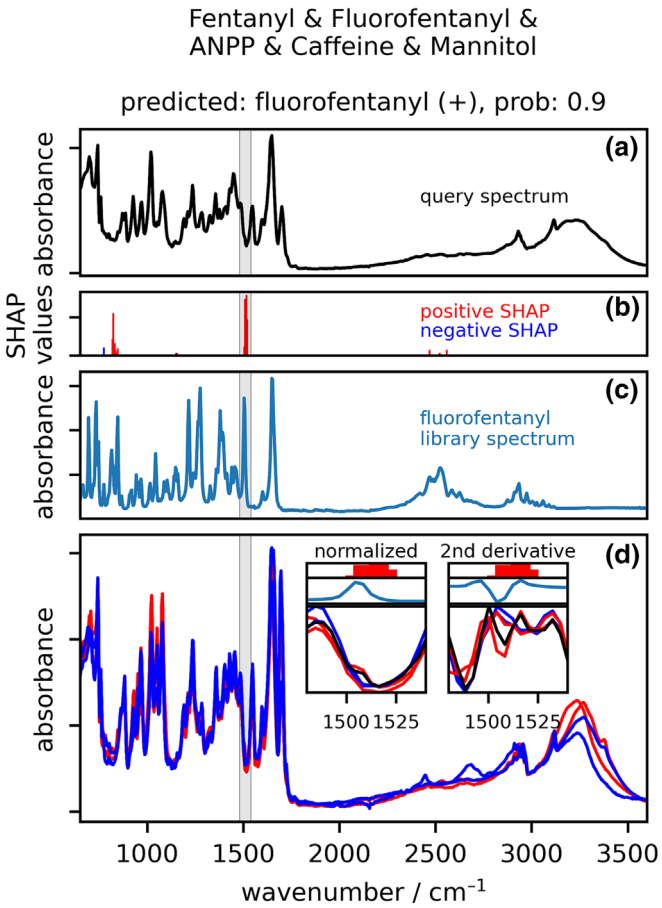

FIGURE 6 Example of a test sample spectrum (a) and explainable AI framework to build adequate trust in the model prediction. (b) SHAP values highlight the spectral regions of interest contributing to the prediction. The (c) library spectrum of fluorofentanyl and (d) the spectra of the 4 nearest neighbors are overlaid for reference. The inset in (d) combines both the *k* nearest neighbors from the positive class (red traces) and negative class (blue traces) and feature highlighting from SHAP in combination with the spectral library of fluorofentanyl to present visual evidence that features consistent with fluorofentanyl exist in the query spectrum (black).

presence of fluorofentanyl with a probability of 90%. Again, the positive SHAP features highlight areas that have contributed to this prediction (red), and features that, according to the trained model, countered this prediction (blue). The features that contributed to the prediction of fluorofentanyl align with characteristic modes from a pure fluorofentanyl spectrum. Figure 6d overlays the positive and negative nearest neighbors, both of which look very similar to the query spectrum. This implies that the features attributed to fluorofentanyl are expected to be subtle relative to features arising from other substance within the mixtures. The most influential region is highlighted on the inset, displayed both with simple min-max normalization and the second derivative preprocessing that was used in model training. This highlights that evidence of fluorofentanyl, though subtle, is present in the query spectrum. In this case, a technician may decide to trust the model prediction based on such evidence.

3.3 | Practical application in harm reduction

In general, implementation of XAI facilitates knowledge production in a way that black-box ML models cannot.[42] It is known that curiosity about drug checking technologies, discussion around drug market trends and expectations, and integrating the personal experience of people is essential to guide drug analysis.[16,17,43] Automating FTIR-based drug checking ultimately aims to extend its reach, particularly to smaller communities and regions lacking public health and harm reduction mandates that may not have the resources to train a technician. One of the main goals of this work was to implement a framework to build adequate trust in an automated model and its predictions. The explanation produced should help present evidence and trust when in fact the prediction is correct, and hesitancy when it is incorrect. Previous studies in explainable AI have found that feature highlighting, as well as presenting explanatory factual and counterfactual cases (as we have done in our various test samples shown in Figure 5 and Figure 6), has assisted users in detecting errors while increasing their understanding of the model itself.[44] This implementation will also contribute to ongoing method re-validation, as the drug supply continues to change. Such potential was demonstrated when the MDA model incorrectly predicted that no MDA was present in an MDA–dimethylsulfone sample.

While ML offers a means to standardize the analysis of IR spectra, harm reduction messaging relies on people who can take those results, assess their validity and provide context in their interpretation. It is always important to consider how incorrect and inconsistent drug checking information might impact interpersonal relationships between the consumer, manufacturer, and distributor of illicit drugs.[17,45,46] Moreover, false positives and negatives affect trust in the service and perceptions of the utility of drug checking.[47–49] This could impact overall engagement with drug checking services, as PWUD often report navigating health and social support services that are stigmatizing, not culturally safe, and not relevant for their needs.[50,51] Ultimately, the optimization and evaluation of ML classification models when used in combination with XAI is poised to even better facilitate a reliable analysis and tailored discussion with service users around their drug checking results.

4 | CONCLUSIONS

Automation has the potential to improve the speed and consistency of drug checking, offering less reliance on technician experience. This work has examined the results of classification models for MDA and fluorofentanyl. In this process, explainable AI was integrated using feature importance via SHAP values and a KNN model to retrieve related/explanatory cases from the training data. The integration of XAI methods provides a level of transparency that facilities continuous knowledge production and engagement of technicians and community members. Such methods will help to bridge the gap between the current role of a drug checking technician and the pursuit of ML methods for drug checking.

ACKNOWLEDGEMENTS

This project was funded by a grant from Health Canada's Substance Use and Addictions Program (SUAP), with additional support from the Vancouver Foundation. High performance computing support and server resource allocation was provided by the University of Victoria. We also acknowledge all the Substance staff and service users as the driving force of this work. Many people have contributed to acquiring and interpreting this dataset. We are extremely grateful for the trust of community members in their donation of drug samples.

ORCID

Chris Gill https://orcid.org/0000-0001-7696-5894
Dennis Hore https://orcid.org/0000-0001-8969-9644

REFERENCES

1. Tupper KW, McCrae K, Garber I, Lysyshyn M, Wood E. Initial results of a drug checking pilot program to detect fentanyl adulteration in a Canadian setting. *Drug Alcohol Depend*. 2018;190:242-245. doi:10.1016/j.drugalcdep.2018.06.020
2. Wallace B, Gozdzialski L, Qbaich A, et al. A distributed model to expand the reach of drug checking. *DHSP*. 2022;23(3):220-231. doi:10.1108/DHS-01-2022-0005
3. Maghsoudi N, Tanguay J, Scarfone K, et al. Drug checking services for people who use drugs: a systematic review. *Addiction*. 2022;117(3):532-544. doi:10.1111/add.15734
4. Masterton W, Falzon D, Burton G, et al. A realist review of how community-based drug checking services could be designed and implemented to promote engagement of people who use drugs. *Int J Environ Res Public Health*. 2022;19(19):11960. doi:10.3390/ijerph191911960
5. Dasgupta N, Figgatt MC. Invited commentary: drug checking for novel insights into the unregulated drug supply. *Am J Epidemiol*. 2022;191(2):248-252. doi:10.1093/aje/kwab233
6. Carroll JJ, Mackin S, Schmidt C, McKenzie M, Green TC. The Bronze Age of drug checking: barriers and facilitators to implementing advanced drug checking amidst police violence and COVID-19. *Harm Reduct J*. 2022;19(1):9. doi:10.1186/s12954-022-00590-z
7. Bardwell G, Boyd J, Tupper KW, Kerr T. "We don't got that kind of time, man. we're trying to get high!": exploring potential use of drug checking technologies among structurally vulnerable people who use drugs. *Int J Drug Policy*. 2019;71:125-132. doi:10.1016/j.drugpo.2019.06.018
8. Harper L, Powell J, Pijl EM. An overview of forensic drug testing methods and their suitability for harm reduction point-of-care services. *Harm Reduct J*. 2017;14(1):52. doi:10.1186/s12954-017-0179-5
9. Meza Ramirez CA, Greenop M, Ashton L, Rehman I. Applications of machine learning in spectroscopy. *Appl Spectrosc Rev*. 2021;56(8-10):733-763. doi:10.1080/05704928.2020.1859525
10. Angulo A, Yang L, Aydil ES, Modestino MA. Machine learning enhanced spectroscopic analysis: towards autonomous chemical mixture characterization for rapid process optimization. *Dig Dis*. 2022;1(1):35-44. doi:10.1039/D1DD00027F
11. Ren H, Li H, Zhang Q, et al. A machine learning vibrational spectroscopy protocol for spectrum prediction and spectrum-based structure recognition. *Fundam Res*. 2021;1(4):488-494. doi:10.1016/j.fmre.2021.05.005
12. Kranenburg RF, Verduin J, Weesepoel Y, et al. Rapid and robust on-scene detection of cocaine in street samples using a handheld near-infrared spectrometer and machine learning algorithms. *Drug Test Anal*. 2020;12(10):1404-1418. doi:10.1002/dta.2895

13. Guo S, Popp J, Bocklitz T. Chemometric analysis in Raman spectroscopy from experimental design to machine learning-based modeling. *Nat Protoc*. 2021;16(12):5426-5459. doi:10.1038/s41596-021-00620-3
14. McCrae K, Tobias S, Grant C, et al. Assessing the limit of detection of Fourier-transform infrared spectroscopy and immunoassay strips for fentanyl in a real-world setting. *Drug Alcohol Rev*. 2020;39(1):98-102. doi:10.1111/dar.13004
15. Ramsay M, Gozdzialski L, Larnder A, Wallace B, Hore DK. Fentanyl quantification using portable infrared absorption spectroscopy. A framework for community drug checking. *Vib Spectrosc*. 2021;114:103243. doi:10.1016/j.vibspec.2021.103243
16. Betsos A, Valleriani J, Boyd J, McNeil R. Beyond co-production: the construction of drug checking knowledge in a Canadian supervised injection facility. *Soc Sci Med*. 2022;314:115229. doi:10.1016/j.socscimed.2022.115229
17. Wallace B, Roode T, Pagan F, Hore D, Pauly B. The potential impacts of community drug checking within the overdose crisis: qualitative study exploring the perspective of prospective service users. *BMC Public Health*. 2021;21(1):1156. doi:10.1186/s12889-021-11243-4
18. Kaluarachchi T, Reis A, Nanayakkara S. A review of recent deep learning approaches in human centered machine learning. *Sensors*. 2021;21(7):2514. doi:10.3390/s21072514
19. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38. doi:10.1016/j.artint.2018.07.007
20. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016; 1135–1144.
21. Yang F, Huang Z, Scholtz J, Arendt DL. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? in Proceedings of the 25th International Conference on Intelligent User Interfaces, Association for Computing Machinery, New York, NY, USA, 2020; 189–201.
22. Cai CJ, Reif E, Hegde N, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. 2019.
23. Gianfagna L. *Explainable AI with Python*. Springer; 2021. doi:10.1007/978-3-030-68640-6.
24. Du M, Liu N, Hu X. Techniques for interpretable machine learning. *Commun ACM*. 2019;63(1):68-77. doi:10.1145/3359786
25. Substance: Vancouver Island Drug Checking Project. https://substance.uvic.ca
26. Borden SA, Saatchi A, Vandergrift GW, Palaty J, Lyshyshyn M, Gill CG. A new quantitative drug checking technology for harm reduction: pilot study in Vancouver, Canada using paper spray mass spectrometry. *Drug Alcohol Rev*. 2022;41(2):410-418. doi:10.1111/dar.13370
27. Borden SA, Saatchi A, Krogh ET, Gill CG. Rapid and quantitative determination of fentanyls and pharmaceuticals from powdered drug samples by paper spray mass spectrometry. *Anal Sci Adv*. 2020;1(2):97-108. doi:10.1002/ansa.202000083
28. Borden SA, Palaty J, Termopoli V, et al. Mass spectrometry analysis of drugs of abuse: challenges and emerging strategies. *Mass Spectrom Rev*. 2020;39(5-6):703-744. doi:10.1002/mas.21624
29. Gozdzialski L, Aasen J, Larnder A, et al. Portable gas chromatography–mass spectrometry in drug checking: detection of carfentanil and etizolam in expected opioid samples. *Int J Drug Policy*. 2021;97:103409. doi:10.1016/j.drugpo.2021.103409
30. Gozdzialski L, Rowley A, Borden S, et al. Rapid and accurate etizolam detection using surface-enhanced Raman spectroscopy for community drug checking. *Int J Drug Policy*. 2022;102:103611. doi:10.1016/j.drugpo.2022.103611
31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
32. Brereton RG. *Applied Chemometrics for scientists*. John Wiley & Sons; 2007. doi:10.1002/9780470057780
33. Lundberg SM, Lee S. A unified approach to interpreting model predictions. *CoRR* 2017; abs/1705.07874.
34. Butler HJ, Smith BR, Fritzsch R, Radhakrishnan P, Palmer DS, Baker MJ. Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy. *Analyst*. 2018;143(24):6121-6134. doi:10.1039/C8AN01384E
35. *Guidance for the validation of analytical methodology and calibration of equipment used for testing of illicit drugs in seized materials and biological specimens*. United Nations Office of Drugs and Crime; 2009.
36. Laing MK, Ti L, Marmel A, et al. An outbreak of novel psychoactive substance benzodiazepines in the unregulated drug supply: preliminary results from a community drug checking program using point-of-care and confirmatory methods. *Int J Drug Policy*. 2021;1:103169. doi:10.1016/j.drugpo.2021.103169
37. Ti L, Tobias S, Lysyshyn M, et al. Detecting fentanyl using point-of-care drug checking technologies: a validation study. *Drug Alcohol Depend*. 2020;212:108006. doi:10.1016/j.drugalcdep.2020.108006
38. Ayres LB, Gomez FJV, Linton JR, Silva MF, Garciia CD. Taking the leap between analytical chemistry and artificial intelligence: a tutorial review. *Anal Chim Acta*. 2021;1161:338403. doi:10.1016/j.aca.2021.338403
39. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
40. Menze B, Kelm B, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinf*. 2009;10(1):213. doi:10.1186/1471-2105-10-213
41. Parmar A, Katariya R, Patel V. A review on random forest: an ensemble classifier. In: Hemanth J, Fernando X, Lafata P, Baig Z, eds. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. Springer; 2019:758-763. doi:10.1007/978-3-030-03146-6_86
42. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inf Fusion*. 2022;77:29-52. doi:10.1016/j.inffus.2021.07.016
43. Barratt MJ, Measham F. What is drug checking, anyway? *DHSP*. 2022;23(3):176-187. doi:10.1108/DHS-01-2022-0007
44. Kenny EM, Ford C, Quinn M, Keane MT. Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error rates in XAI user studies. *Artif Intell*. 2021;294:103459. doi:10.1016/j.artint.2021.103459
45. Bardwell G, Boyd J, Arredondo J, McNeil R, Kerr T. Trusting the source: the potential role of drug dealers in reducing drug-related harms via drug checking. *Drug Alcohol Depend*. 2019;198:1-6. doi:10.1016/j.drugalcdep.2019.01.035
46. Carroll JJ. Auras of detection: power and knowledge in drug prohibition. *Contemp Drug Probl*. 2021;48(4):327-345. doi:10.1177/00914509211035487
47. Betzler F, Helbig J, Viohl L, et al. Drug checking and its potential impact on substance use. *Eur Addict Res*. 2021;27(1):25-32. doi:10.1159/000507049
48. Mema SC, Sage C, Xu Y, et al. Drug checking at an electronic dance music festival during the public health overdose emergency in British Columbia. *Can J Public Health*. 2018;109(5-6):740-744. doi:10.17269/s41997-018-0126-6
49. Wallace B, Roode T, Pagan F, et al. What is needed for implementing drug checking services in the context of the overdose crisis? A qualitative study to explore perspectives of potential service users. *Harm Reduct J*. 2020;17(1):29. doi:10.1186/s12954-020-00373-4
50. Murney MA, Sapag JC, Bobbili SJ, Khenti A. Stigma and discrimination related to mental health and substance use issues in primary health care in Toronto, Canada: a qualitative study. *Int J Qual Stud*

Health Well-Being. 2020;15(1):1744926. doi:10.1080/17482631.2020.1744926

51. Neufeld SD, Chapman J, Crier N, Marsh S, McLeod J, Deane LA. Research 101: a process for developing local guidelines for ethical research in heavily researched communities. *Harm Reduct J*. 2019;16(1):41. doi:10.1186/s12954-019-0315-5

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gozdzialski L, Hutchison A, Wallace B, Gill C, Hore D. Toward automated infrared spectral analysis in community drug checking. *Drug Test Anal*. 2024;16(1):83-92. doi:10.1002/dta.3520